Unveiling Tissue Structure and Tumor Microenvironment from Spatially Resolved Transcriptomics by Hypergraph Learning

Yi Liao[1], Chong Zhang[1], Zhikang Wang[2], Fei Qi[1], Weitian Huang[1], Shangyan Cai[3], Junyu Li[1], Zhiyuan Yuan[4,5,*], Jiangning Song[2,*], Hongmin Cai[1,*]

[1] School of Computer Science and Engineering, South China University of Technology, Guangzhou, Guangdong, 510006, China;
[2] Biomedicine Discovery Institute and Department of Biochemistry and Molecular Biology, Monash University, Melbourne, VIC, 3800, Australia;
[3] Faculty of Science and Technology, Beijing Normal University-Hong Kong Baptist University United International College, Zhuhai, Guangdong, 519087, China;
[4] Institute of Science and Technology for Brain-Inspired Intelligence, MOE Key Laboratory of Computational Neuroscience and Brain-Inspired Intelligence, MOE Frontiers Center for Brain Science, Fudan University, Shanghai, 200433, China;
[5] Center for Medical Research and Innovation, Shanghai Pudong Hospital, Fudan University Pudong Medical Center, Fudan University, Shanghai, 201399, China.

* To whom correspondence should be addressed. Hongmin Cai: hmcai@scut.edu.cn; Jiangning Song: Jiangning.Song@monash.edu; Zhiyuan Yuan: zhiyuan@fudan.edu.cn.

Abstract

Spatially resolved transcriptomics (SRT) technologies acquire gene expressions and spatial information simultaneously, reshaping the perspectives of life sciences. Identifying spatial patterns is essential for exploring organ development and tumor microenvironment. Nevertheless, emerging SRT technologies have also introduced diverse spatial resolutions, posing challenges in characterizing spatial domains with finer resolutions. Here we propose a hypergraph-based method, termed HyperSTAR to precisely recognize spatial domains across varying spatial resolutions by utilizing higher-order relationships among spatially adjacent tissue programs. Specifically, a gene expression-guided hyperedge decomposition module is incorporated to refine the structure of the hypergraph to precisely delineate the boundaries of spatial domains. A hypergraph attention convolutional neural network is designed to adaptively learn the significance of each hyperedge. With the power of capturing intricate higher-order relationships within spatially neighboring multi-spots/cells, HyperSTAR demonstrates superior performance across different technologies with various resolutions compared to existing advanced graph neural network models in multiple tasks including uncovering tissue sub-structure, inferring spatiotemporal patterns, and denoising spatially resolved gene expressions. It successfully reveals spatial heterogeneity in breast cancer section and its findings are further validated through functional and survival analyses of independent clinical data. Notably, HyperSTAR performs well with diverse spatial omics data types and seamlessly extends to large-scale datasets.

Introduction

In living organisms, biological process happens within a spatial context. The location of each cell is as crucial as its inherent characteristics in determining cell types, cell states and functions [1]. Single-cell omics quantifies genomes, epigenomes, transcriptomes and proteomes at the single-cell level [2-3]. Nevertheless, the process of single-cell dissociation during sample preparation results in the loss of spatial information [4]. Emerging spatially resolved transcriptomics (SRT) have facilitated the profiling of gene expression while preserving spatial location in tissues, offering valuable opportunities in exploring intricate mechanisms within biological systems [5].

Widely used SRT can be broadly categorized into two groups: (1) imaging-based approaches, and (2) next-generation sequencing (NGS)-based approaches [6-10]. These diverse technologies simultaneously capture both the gene expression and spatial location, while differing from gene throughput, sensitivity, resolution, and field of view (fov). Imaging-based methods provide subcellular resolution with high sensitivity but limited gene throughout [11-17], while NGS-based technologies generate diverse spatial resolutions, ranging from multi-cellular to subcellular levels for the captured spots with whole-transcriptome coverage but lower sensitivity [18-20].

Several computational methods have been developed to identify spatial domains in SRT data. Earlier works utilized traditional methods (e.g. K-means [21-22], Louvain [23]), which did not take into account the spatial information of SRT data, leading to a discontinuous division of regions. Recent methods have introduced spatial information to enhance spatial domain identification. For example, BayesSpace [24] employed a Bayesian statistical approach, using the information from spatial neighborhoods for spatial clustering. Deep learning methods have also been developed to integrate spatial information and gene expression of SRT data. StLearn [25] defined morphological distances based on features extracted from histology images and utilized these distances, along with spatial neighbors to smooth gene expressions. SpaGCN [26] also used histology images by employing graph convolutional networks to integrate gene expression, spatial location and image. SpaceFlow [27] utilized spatially regularized deep graph networks to integrate both expression similarity and spatial information. STAGATE [28] applied a graph attention auto-encoder to learn low-dimensional latent embeddings of SRT profiles. Although these methods have shown significant improvements on some datasets, they primarily focus on the cell-cell pairwise relationships, neglecting the higher-order structures inherent in multicellular programs. Furthermore, they fail to account for the delineation of domain boundaries, a crucial aspect for enhancing spatial domain results, as they predominantly focused on clustering. This limitation becomes more pronounced when

confronted with higher spatial resolutions, making it challenging to elucidate the tissue structure comprehensively. Moreover, these approaches encounter difficulties in extending their capabilities to accommodate other omics.

To tackle these issues, we propose HyperSTAR, the first hypergraph learning-based spatial clustering method by explicitly learning the higher order tissue structures. HyperSTAR can effectively leverage complex relationships among spatially adjacent multi-cellular programs. A gene expression-guided hyperedge decomposition module and a hypergraph attention convolutional layer are employed to precisely delineate boundaries of spatial domains. Extensive experiments show HyperSTAR could reveal tissue structure and tumor spatial heterogeneity across varying spatial resolutions, demonstrating effectiveness at diverse downstream analyses, such as spatiotemporal trajectories inference, gene expression denoising and spatially variable genes (SVGs) detection.

Results

Overview of HyperSTAR

Spatial resolved transcriptomics (SRT) data typically contains a gene expression matrix and a spatial coordinate matrix, as illustrated in Fig. 1a. To comprehensively capture the intricate associations among spatially adjacent multiple spots, we develop a hypergraph-based method, as illustrated in Fig. 1b and Fig. 1c. The construction of this hypergraph can be primarily divided into two key steps: (1) acquiring spot representations and (2) establishing the hypergraph structure. Spot representations are acquired from pre-processed gene expression data. The hypergraph structure is attained through two stages: (1) constructing a naive hyperedge set among spatially neighboring multi-spots, and (2) conducting a gene expression-guided hyperedge decomposition module to refine these naive hyperedges. Finally, a hypergraph structure that fuses spatial information and gene expression is obtained (Fig. 1b). The initially built naive hyperedge set contains complex high-order relationships among spatially neighboring multi-spots. However, it is important to note that adjacent spots in space do not always belong to the same domain. For instance, although spots near a domain boundary are spatially adjacent, they might actually belong to different domains. To tackle this issue, a gene expression-guided hyperedge decomposition module is conducted to enhance the effects of gene expression. After constructing the hypergraph, we utilize hypergraph attention convolutional neural networks to learn low-dimensional latent embeddings of the SRT data. Subsequently, the latent embeddings are used to visualize data, identify spatial domains, infer trajectories, detect spatially variable genes (SVGs), and denoise spatial gene expression (Fig. 1c). Using the above strategy to integrate spatial information and

gene expression in SRT data, HyperSTAR improves the performance of spatial omics analysis.

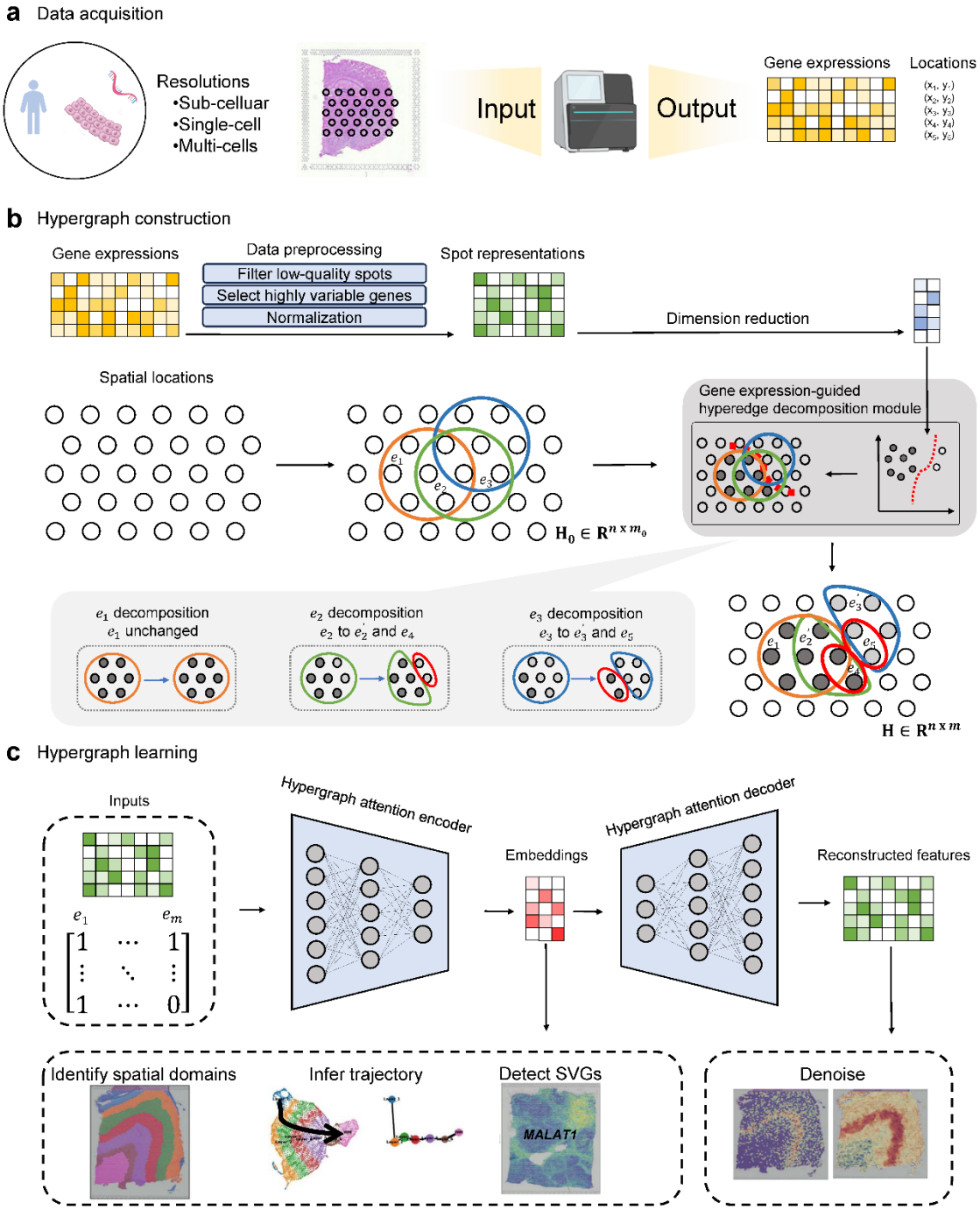

Fig. 1 Overview of HyperSTAR. a Spatial resolved transcriptomics (SRT) data varies in resolution depending on the sequencing technologies. Generally, SRT data consists of gene expression and corresponding spatial locations. **b** A hypergraph is first constructed to model the relationships among multi-spots. Spot representations are directly derived from gene expression data. The hypergraph structure is constructed in two stages: first, for each spot, a hyperedge consisting of its spatially adjacent multi-spots is formed. Subsequently, hyperedge decomposition is performed to refine the hyperedges according to the clustering results of gene expression data. **c** The final refined hyperedge set and spot representations are fed into hypergraph attention convolutional neural networks to learn latent embeddings of spots for a series of downstream analyses.

HyperSTAR achieves state-of-the-art performance on the benchmark dataset

To evaluate the performance of HyperSTAR, we first adopted the widely used SpatialLIBD dataset, which is a 10x Visium dataset of 12 human dorsolateral prefrontal cortex (DLPFC) sections [29]. SpatialLIBD has manually annotated DLPFC layers (Layer 1 to Layer 6) and white matter (WM) based on morphological features and marker genes. Regarding these annotations as the ground truth, we employ the clustering metric Adjusted Rand Index (ARI) score to quantitatively assess the performance of spatial domain identification.

We compare our method HyperSTAR with six commonly used packages for identifying spatial domains: Scanpy [23] (originally designed for single-cell omics), two methods that integrate histology images--stLearn and SpaGCN, two deep learning methods based on graphs--SpaceFlow and STAGATE, and one Bayesian statistical-based method, BayesSpace. HyperSTAR effectively identified the expected cortical layer structures, demonstrating significant improvement compared to other methods (Fig. 2a). For example, in sections 151673 and 151507 (ground truth shown in Fig. 2b), HyperSTAR clearly delineated the layer borders and achieved the highest clustering accuracy (ARI = 0.57 and ARI = 0.60, respectively) (Fig. 2c, 2d). The results of spatial visualization for all sections are provided in Supplementary Fig. S1-S15. We further confirmed the inferred trajectory using a trajectory inference algorithm named PAGA. The PAGA graphs illustrate that embeddings belonging to different clusters obtained by Scanpy, stLearn, and SpaceFlow are intermingled. As for STAGATE, it successfully separates WM, Layer 6, and Layer 5 clearly; however, Layer 1, Layer 2, Layer 3, and Layer 4 remain mixed. Our HyperSTAR embeddings effectively divide these cortical layers, providing a clear depiction of tissue topology (Fig. 2e).

We further conducted ablation experiments to illustrate the effectiveness of our strategies. In Fig. 2f, we present the performance of four graph-based networks: graph convolution networks (GCN) [30], graph attention networks (GAT) [31], hypergraph convolution networks (HGNN) [32], and our HyperSTAR, which combines hypergraph convolution operation and a graph attention mechanism. As shown in Fig. 2f, the ARI score of HyperSTAR outperforms other methods on almost all slices of the DLPFC dataset. In Fig. 2g, UMAP plots of embeddings obtained from four networks demonstrate the superiority of HyperSTAR. For GCN, GAT, and HGNN, there is some overlap between cortical layers, indicating some spots are mixed. In contrast, HyperSTAR clearly separates all cortical layers without any mixing.

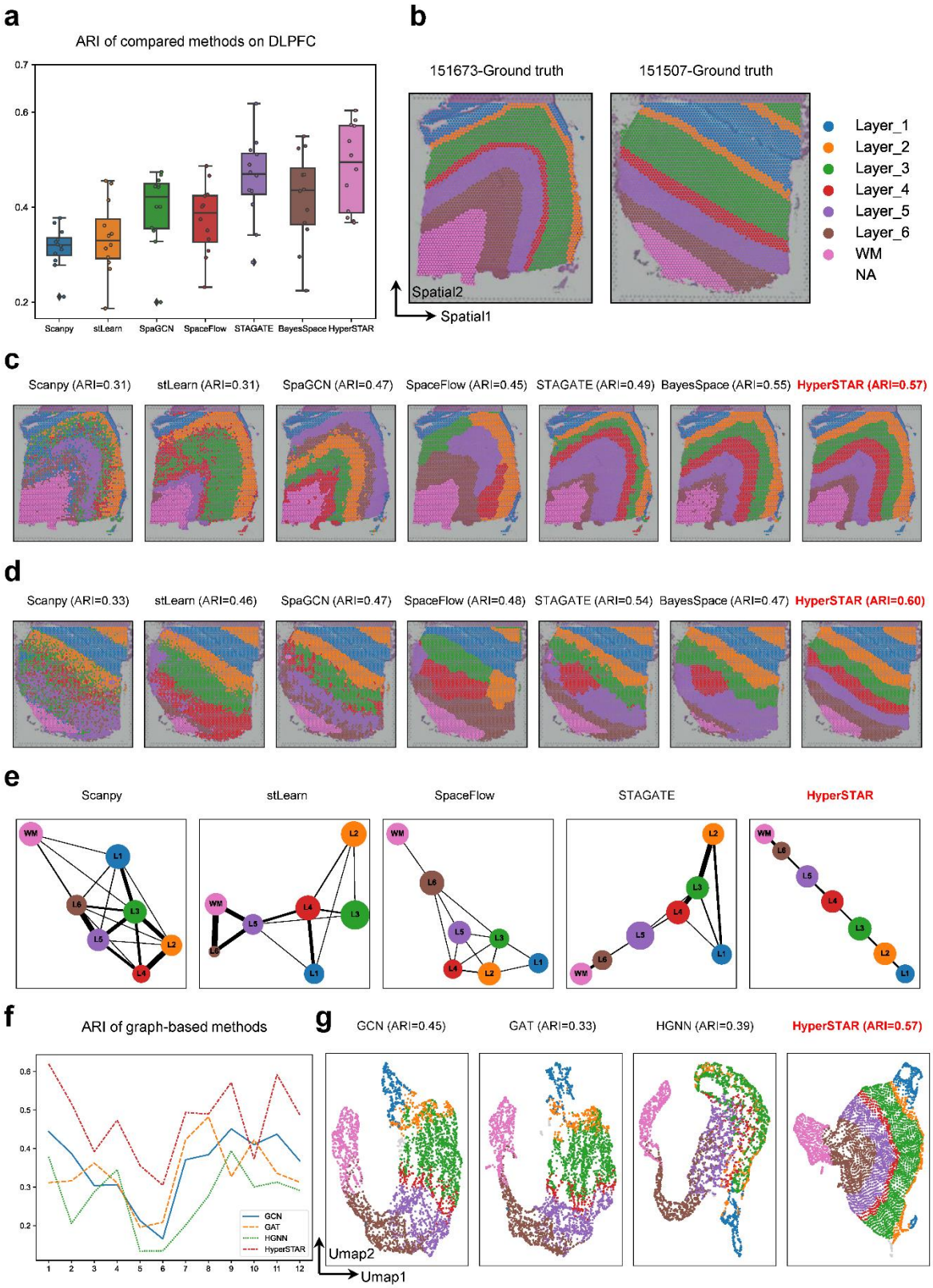

Fig. 2 HyperSTAR achieves state-of-the-art performance on the DLPFC benchmark dataset. a Boxplots of clustering accuracy across all 12 sections of the dataset in terms of ARI scores for seven methods. **b** Ground-truth segmentation of cortical layers (from WM to Layer 1) and white matter in slices 151673 and 151507. **c** Spatial visualization of clustering results in slice 151673 compared to six state-of-the-art methods. **d** Spatial visualization of clustering results in slice 151507 compared to six state-

of-the-art methods. **e** The PAGA graphs generated by Scanpy, stLearn, SpaceFlow, STAGATE, and HyperSTAR. **f** Line plots of clustering accuracy across all 12 sections of the dataset in terms of ARI scores for four different graph-based networks (including GCN, GAT, HGNN, and HyperSTAR). **g** The UMAP plots of section 151673 obtained by the four graph-based networks.

HyperSTAR reveals sub-structure on finer resolutions of mouse olfactory bulb

To assess HyperSTAR's performance across different resolutions, we selected a mouse olfactory bulb section from Stereo-seq, with finer spatial resolution at the single-cell level. Spatial domains on this section were initially identified using three methods: Scanpy, a widely used single-cell method; STAGATE, a graph-based deep learning method; and our proposed method, HyperSTAR. As depicted in Fig. 3b, Scanpy yielded discontinuous results, while STAGATE showed improved performance, albeit with some mixing between layers, such as the fusion of domain 3 and domain 4. In contrast, HyperSTAR identified a clear and continuous layer structure in the mouse olfactory bulb, corresponding to the mouse anatomy collected from the Allen Brain Atlas, as shown in Fig. 3a.

To further analyze the results obtained by HyperSTAR, we ranked genes for characterizing groups, and the top genes identified for each group are shown in Fig. 3e. The gene expressions of *Calb2*, *Pcp4*, and *Ptgds*, shown in Fig. 3d, represent the three top genes identified by HyperSTAR for characterizing domains 7, 4, and 2, respectively. The distribution of domains 7, 4, and 2 identified by HyperSTAR is displayed in Fig. 3c, corresponding to the reference in Fig. 3a as layers OPL, GR, and GL. The visualization of other marker genes distinguishing the four layers is also provided in Supplementary Fig. S16, collectively illustrating HyperSTAR's ability to reveal continuous tissue structure and find effective markers.

We adjusted the parameters for fine-grained clustering, and the result is shown in Fig. 3g. HyperSTAR reveals a sub-structure ACO within GR (Fig. 3h) and its top characterizing gene Mbp. Additionally, we identified the top gene in GR as Xrxn3. The In Situ Hybridization (ISH) image illustrates the effectiveness of the result obtained by HyperSTAR. The heatmap of the correlation matrix of clustering also demonstrates the effectiveness of HyperSTAR, showing similarity within clusters and dissimilarity between clusters (Fig. 3f).

We further assessed HyperSTAR's cross-resolution capability on various spatial resolutions of the above mouse olfactory bulb section. Across bin 20, bin 50, bin 100, and single-cell levels (Supplementary Fig. S17 and Fig. 3b), our HyperSTAR results consistently exhibit a continuous structure aligned with the reference in Fig. 3a. Extending our evaluation to an adult axolotl brain slice further confirmed HyperSTAR's efficacy, revealing a continuous structure corresponding to the reference (Supplementary Fig. S18).

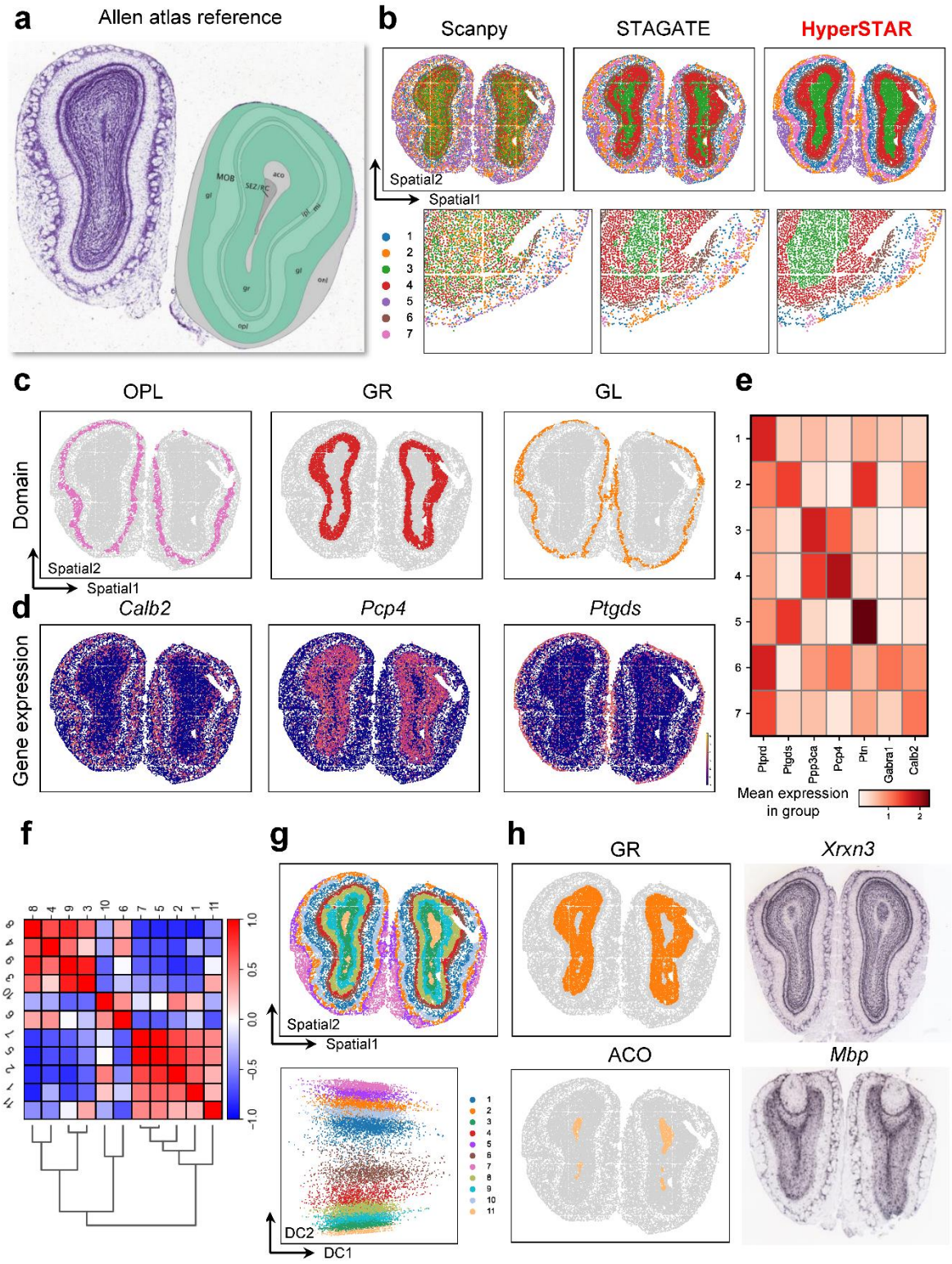

Fig. 3 HyperSTAR reveals sub-structure ACO on finer resolutions of the Stereo-seq mouse olfactory bulb. a Mouse brain olfactory bulb anatomical structure from the Allen Brain Atlas. **b** Spatial visualization of results obtained by three spatial domain identification methods: Scanpy, STAGATE, and HyperSTAR. **c** OPL, GL, and GR layers identified by HyperSTAR. **d** The spatial visualization of three marker genes expressing in correspondence to OPL (*Calb2*), GL (*Pcp4*), and GR (*Ptgds*), respectively. **e** Heatmap displays top genes characterizing seven groups identified by HyperSTAR. **f** Correlation matrix of results obtained from HyperSTAR, showing similarity within clusters and dissimilarity between

clusters. **g** Top to bottom: spatial visualization results from HyperSTAR and diffusion map plot obtained from HyperSTAR, revealing the sub-structure ACO within GR. **h** Top to bottom: GR identified by HyperSTAR and ISH image of corresponding marker gene *Nrxn3* collected from the Allen Brain Atlas; ACO identified by HyperSTAR and ISH image of corresponding marker gene *Mbp* collected from the Allen Brain Atlas.

HyperSTAR unveils spatial intra-tumor heterogeneity in breast cancer section and its association with survival

To investigate the capability of HyperSTAR for tumor analysis, we applied HyperSTAR on a breast cancer section obtained from 10x Visium. The section was manually divided into four categories by pathologist: ductal carcinoma in situ/lobular carcinoma in situ (DCIS/LCIS), invasive ductal carcinoma (IDC), healthy tissue, and the tumor edge (Fig. 4a) [33]. HyperSTAR accurately identified these categories and showing strong consistency with the manual annotation (Fig. 4b).

We next interrogate the diagnosis power of HyperSTAR's results. We rank and plot top 5 genes for characterizing these 4 domains (Fig. 4c). For *KRT8* and *KRT18*, which are found highly expressed in IDC region. *KRT8* is an independent prognostic indicator for poor survival in lung adenocarcinoma tissue [34]. *KRT18* is often together with *KRT8*, involved in interleukin-6 (IL-6)-mediated barrier protection. This patient has been clinically diagnosed as Grade III. Logan and et al. claims that *KRT8/18* expression differentially distinct subtypes of grade III invasive ductal carcinoma of the breast cancer [35], which is consistent with our results. We further compared transcriptional differences between IDC and the rest by performing differential expression analysis followed by pathway enrichment analysis. The differential expression genes (DEGs) of IDC are enriched in the process of mitosis, indicating abnormal proliferation of cancer cells (Fig. 4d), which indicates the area is probably infiltrated with cancer cells and also highly active. We further investigate the correlation of these DEGs with survival on an independent clinical data from TCGA breast cancer cohort. 10 signature genes (Supplementary Table 1) were selected to cutoff patients, and the result shows that low expression of these 10 signatures correlated with shorter overall survival (Fig. 4e). Hence, these results indicate that there is possible cell development from cancer stem cells to malignant cells. Here are also some research claims that *KRT8* is a pan-cancer indicator [36], In light of this, we conducted a pan-cancer survival analysis using these DEGs identified in IDC. The outcomes further reveal a correlation between these DEGs and pan-cancer development (Suppementary Fig. S19).

We further investigate the subtypes of IDC region. From our HyperSTAR, the IDC is divided into 4 subtypes (Fig. 4f). The heatmap, generated by the top two genes of each subtype (Figure 4g), highlights the unique expression profiles within each subtype. Notably, for Domain 4, *SLTRK6* and *CRISP3* exhibit high expression levels, with *SLITRK6* emerging as a potential

marker for this specific subtype. The accompanying boxplot illustrates that *SLTRK6* is significantly upregulated in IDC_SLTRK6 compared to IDC, tumor, and non-tumor regions (Figure 4i). In summary, HyperSTAR not only enables the identification of intra-tumor heterogeneity but also facilitates the exploration of distinct subtypes within the intra-tumoral environment.

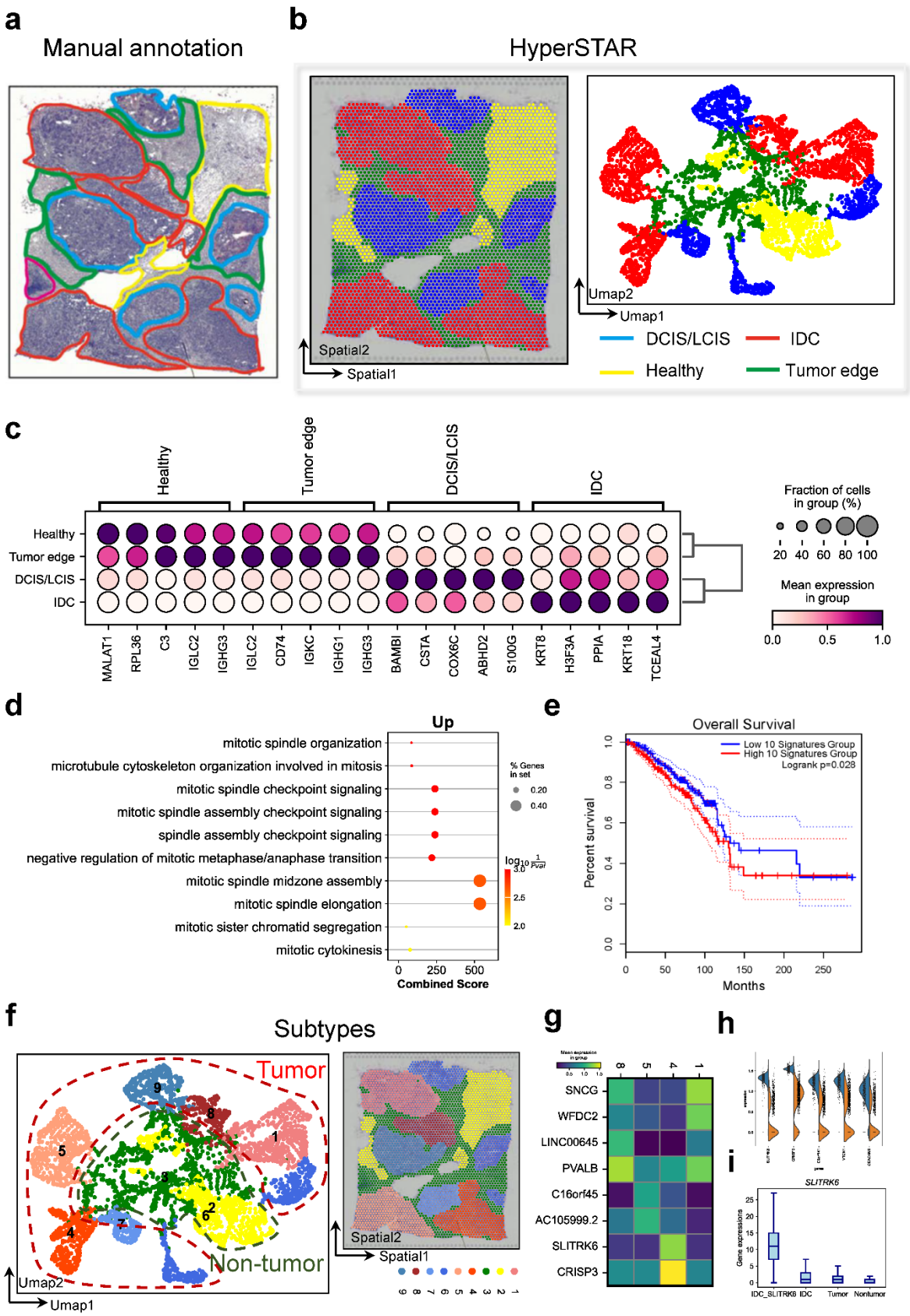

Fig. 4 HyerSTAR identifies gene markers in subtypes of breast cancer and association with survival. a Manual annotation of breast cancer section from 10x Visium by pathology experts. **b** Spatial

results and UMAP plots by HyperSTAR on this section. **c** Dotplots of marker genes identified by HyperSTAR. **d** Gene ontology enrichment analysis of the DEGs between IDC and rest. **e** Total survival rate of patients of 10 signature genes for IDC in gene expression data of breast cancer from TCGA by GEPIA2. **f** Subtypes identified by HyperSTAR in breast cancer section. **g** Heatmaps of marker genes in subtypes of IDC. **h** Top five marker genes of subtypes IDC_SLITRK6 between IDC_SLITRK6 and the rest of IDC. **i** Gene expressions of *SLITRK6* in IDC_SLITRK6, IDC, tumor and non-tumor.

HyperSTAR is generalisable

To validate the capability of HyperSTAR on various technologies and different omics, we applied it to multiple platforms including 10x Genomics (10x Visium, 10x Xenium), MERFISH [37], NanoString (CosMx), and osmFISH (Suppementary Fig. S20, Fig. 5). Our evaluation also spanned different omics layers, encompassing both transcriptomics and proteomics.

HyperSTAR reveals the cortex distribution of the somatosensory cortex (Fig. 5a) in a dataset produced by osmFISH, a technique based on multiplexed fluorescence *in situ* hybridization [38]. osmFISH enables the detection of hundreds of genes at single-cell resolution, providing valuable insights into the spatial organization of complex tissues. Applying HyperSTAR to this dataset, we were able to identify distinct layers and regions within the somatosensory cortex, showcasing the method's ability to uncover biologically meaningful structures. To further demonstrate the versatility of HyperSTAR, we applied it to the STARmap [39] mouse brain data. As a result, HyperSTAR explicitly revealed the distribution of brain layers (Fig. 5c, 5d), highlighting its effectiveness in identifying spatial patterns across different experimental platforms. The heatmaps of the result in STARmap data display the correlation matrix of the identified clusters, with high correlation values within clusters but low correlation values between clusters (Fig. 5b). This indicates that HyperSTAR successfully captures the underlying structure of the data, grouping together domains with similar spatial gene expression profiles while separating domains from different layers or regions.

Hypergraph is an easily scalable framework which can be applied to jointly cluster data from multiple sources. To assess its capability, we chose mouse olfactory bulb SRT data from SlideSeqV2 and Stereo-seq, two different sequencing platforms. To integrate the data from these different sequencing platforms, we extended the hypergraph construction process beyond a single slice. In addition to building hyperedges within each slice, we established relationships among multiple slices by constructing hyperedges that connect spots across the slices based on their normalized log-transformed gene expression profiles. This approach enables the identification of shared spatial patterns across the different datasets. Furthermore, to mitigate potential batch effects arising from the different sequencing platforms, we employed Harmony. The integration results (Fig. 5g, 5h) demonstrate the successful removal of batch effects, allowing for a more accurate joint clustering analysis. The joint clustering results obtained by HyperSTAR (Fig. 5e, 5f) reveal the presence of distinct layers in the mouse

olfactory bulb, consistent with the known anatomical structure of this region. Taken together, by leveraging the complementary information provided by SlideSeqV2 and Stereo-seq, HyperSTAR can identify shared spatial patterns across the datasets, providing a more comprehensive understanding of the tissue organization.

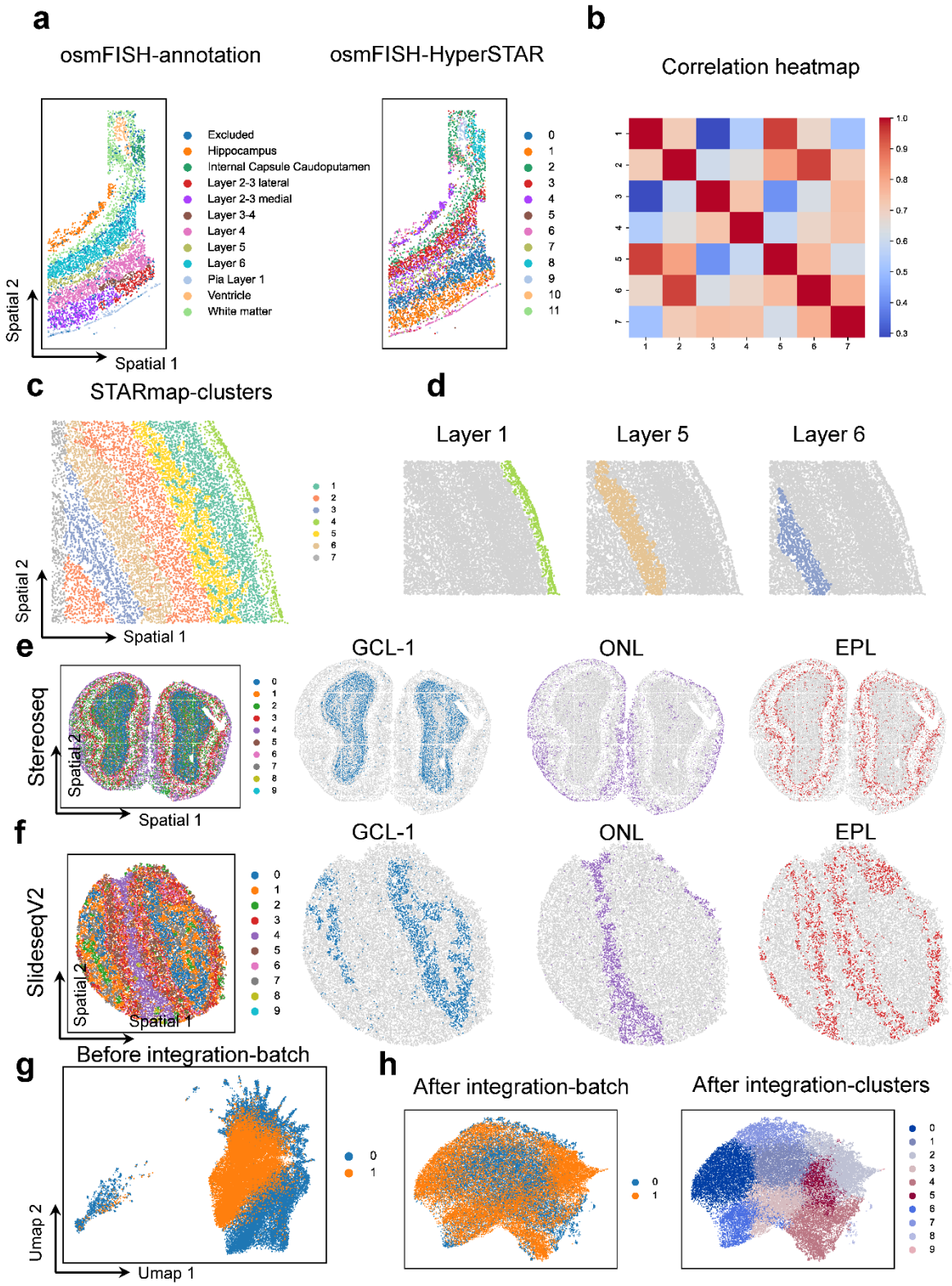

Fig. 5 HyperSTAR is applicable to various spatial omics datasets with demonstrated ability of joint clustering. a Manual annotation and HyperSTAR identification of the somatosensory cortex from osmFISH data. **b** Correlation heatmaps of the clustering results obtained by HyperSTAR on STARmap data. **c** Spatial results of HyperSTAR on a mouse brain section from STARmap. **d** Spatial distribution of different layers revealed by HyperSTAR on the STARmap data. **e** Joint clustering results of Stereo-seq data and spatial distribution of different layers identified by HyperSTAR. **f** Joint clustering results of SlideSeqV2 data and spatial distribution of different layers identified by HyperSTAR. **g** UMAP visualization of the slices from Stereo-seq and SlideSeqV2 prior to data integration. **h** UMAP plots of the SlideSeqV2 and Stereo-seq slices after the HyperSTAR integration.

Discussion

The development of spatial omics technologies has revolutionized our understanding of biological processes by enabling the simultaneous acquisition of biomolecular and spatial information. Accurate identification of spatial domains is critical for deciphering spatiotemporal life processes. Here, we have developed a powerful method for spatial omics analysis called HyperSTAR, which can accurately identify spatial domains, even with high-resolution spatial data. Through the incorporation of a gene expression-guided hyperedge decomposition module and a hypergraph attention convolutional neural network, HyperSTAR excels in precisely recognizing spatial domains across varying resolutions. As such, the findings here have important implications for the exploration of organ development and tumor microenvironments.

Benchmarking and comparison of HyperSTAR with existing advanced graph neural network models highlights its superior performance in uncovering tissue sub-structures, inferring spatiotemporal patterns, and denoising spatially resolved gene expressions. The application of HyperSTAR to spatial omics data from different platforms, including 10x Genomics, NanoString, CODEX, and osmFISH, demonstrates its adaptability and robustness across various technologies. Furthermore, the successful validation of HyperSTAR's findings through functional and survival analyses of independent clinical data, particularly in the context of breast cancer, adds credibility to its utility in deciphering spatial heterogeneity.

The strength of HyperSTAR lies in its ability to capture intricate higher-order relationships within spatially neighboring multi-spots. Nevertheless, the current version of HyperSTAR is mainly used for identifying spatial domains and does not fully exploit the full potential of hypergraphs. In future work, we plan to further extend HyperSTAR to integrate multi-omics data from spatial omics experiments generated by different platform techniques. Additionally, incorporating histology images can be a valuable strategy for addressing clinically revelant questions and further enhancing the method's applicability.

In conclusion, HyperSTAR represents a significant advancement in spatial omics analysis, overcoming challenges associated with diverse spatial resolutions. Its excellent ability to

capture intricate higher-order relationships within spatially neighboring multi-spots makes it a valuable tool for researchers to explore and characterize key spatial patterns in organ development and tumor microenvironments. This study establishes HyperSTAR as a versatile and effective computational approach with potential applications in diverse spatial omics datasets and large-scale analyses. Future work may focus on further refining and expanding the capabilities of HyperSTAR, as well as exploring its application in additional biological contexts and datasets. Integrating multi-omics data and histology images could be promising avenues for future improvements and extensions of our approach.

Methods

Data description

We applied HyperSTAR to various spatial omics datasets generated by different technologies to demonstrate the effectiveness of our model. These technologies include SRT platforms such as 10x Visium, Stereo-seq, osmFISH, MERFISH and Xenium, as well as CODEX, which is used for generating spatial proteomics data. These technologies generate data with varying resolutions, ranging from spot resolution to single-cell resolution. Datasets from 10x Visium have a resolution of 55 μm (1~10 cells per spot), while Stereo-seq provides spot resolutions with bin sizes of 20, 50, 100, and single-cell resolution. Specifically, we used the human dorsolateral prefrontal cortex (DLPFC) dataset [29], generated by 10x Visium, as the benchmark dataset for spatial domain identification. This dataset contains 12 human DLPFC sections sampled from three individuals, with manual annotations provided by the original authors. The 10x Visium mouse brain coronal dataset comprises 2,702 spots and 21,949 genes as the number of raw features. The Stereo-seq mouse brain olfactory bulb dataset, with single-cell resolution, includes 19,109 spots and 27,106 genes as the number of raw features. The breast cancer section dataset, generated by 10x Visium, has a median of 20,762 unique molecular identifiers (UMI) counts per spot and a median of 6,026 genes per spot. The datasets we analyzed are diverse, covering different species, organs, and resolutions.

Data preprocessing

To effectively explore the raw data, it is essential to employ appropriate preprocessing steps. Initially, we remove spots located outside the primary tissue region to focus on the biologically relevant areas. Subsequently, a feature selection process is performed to obtain the top highly variable genes (HVGs) (normally we used 3,000). After selecting the HVGs, we apply a log transformation to the gene expression values. The resulting normalized HVGs are then utilized as input node representations in the hypergraph learning module. By using these preprocessed and normalized gene expression values, the hypergraph learning module can

more effectively capture the intricate relationships among the spots and identify biologically meaningful spatial domains.

Hypergraph construction

After data preprocessing to obtain spot representations, the next step is to construct the hypergraph structure, which involves two steps: Firstly, for each spot, we generate a hyperedge using the k-nearest neighbors (kNN) approach according to the spatial information of the spots. Such initial hyperedge construction captures the local spatial relationships among the spots, ensuring that spatially adjacent spots are connected in the hypergraph. However, considering that the initial hyperedge construction might not handle the scenario where spots are spatially adjacent but belong to different domains, we refine the hyperedge set by utilizing pre-clustering results based on gene expression. This refinement step helps to align the hypergraph structure with the underlying biological domains, thereby improving the accuracy of the spatial domain identification. Specifically, if the spots within a hyperedge belong to the same cluster based on the pre-clustering results, the hyperedge remains unchanged. This indicates that the spatially adjacent spots also share similar gene expression patterns, suggesting they likely belong to the same biological domain. However, if the spots within a hyperedge belong to different clusters, there exist two possible scenarios to consider: (1) When only one spot belongs to a different cluster, we remove that particular spot from the hyperedge. This ensures that the hyperedge only contains spots that are both spatially adjacent and share similar gene expression patterns; (2) When multiple spots belong to different clusters, we decompose the hyperedge into two new hyperedges. Each new hyperedge will contain spots that belong to the same cluster, effectively separating the spots that are spatially adjacent but have distinct gene expression patterns. By refining the hyperedge set based on the pre-clustering results, we can better capture the spatial and biological relationships among the spots, leading to a more accurate representation of the spatial domains in the hypergraph structure.

Hypergraph learning

Once the hypergraph is constructed, we input the spot features and hypergraph structure into the hypergraph learning module, which consists of two parts: the encoder and the decoder. The encoder generates latent low-dimensional embeddings of spots used for identifying spatial domains and downstream analyses. The decoder produces reconstructed features of spots utilized for denoising.

Given a hypergraph **G** = (**X**, **H**), where $\mathbf{G} \in \mathbf{R}^{n\times d}$ represents the feature matrix of *n* spots and $\mathbf{H} \in \mathbf{R}^{n\times m}$ represents incidence matrix from vertexes to hyperedges defined as [32, 40]:

$$h_{ij} = \begin{cases} 1, & \text{if } x_i \in N_j^e \\ 0, & \text{otherwise} \end{cases},$$

where *N* denotes the set of vertexes connected to the *j*-th hyperedge.

Encoder

The encoder takes the normalized gene expressions $\mathbf{X}_0$ as input and generates the spot embeddings by collectively aggregating the information from its neighbors in the same hyperedge. To adaptively learn the weights of the hyperedges, we used three layers of hypergraph attentional layers.

Hypergraph convolutional layer

Message propagation on the hypergraph is more complex than the graph, involving two stages: attentive hyperedge aggregation and attentive vertex aggregation [41]. In the attentive hyperedge aggregation, the connected vertexes in the same hyperedge are aggregated to generate the hyperedge embeddings. As for the *t*-th layer, firstly, the hyperedge features **E** can be calculated by

$$\mathbf{E}_{t-1} = \mathbf{H}^T \mathbf{X}_{t-1},$$

where $\mathbf{X}_{t-1}$ represents output vertex features of the last layer. For each pair of (e_i, e_j) ($e_j \in N_i^v$), we calculate the coefficient a_{ij} between the hyperedge i and hyperedge j by introducing the attention mechanism [42]:

$$a_{ij} = \frac{\exp(\text{LeakyReLU}(\vec{\Theta}_a \left[\mathbf{W}_a \vec{e_i} \,||\, \mathbf{W}_a \vec{e_j} \right]))}{\sum_{e_k \in N_i^v} \exp(\text{LeakyReLU}(\vec{\Theta}_a \left[\mathbf{W}_a \vec{e_i} \,||\, \mathbf{W}_a \vec{e_k} \right]))},$$

where $\mathbf{W} \in \mathbf{R}^{d' \times d}$ is a shared linear transformation, parametrized by a weight matrix. $\vec{\theta}_a \in R^{2d'}$ represents a shared attentional mechanism, parametrized by a vector. N_i^v denotes the set of hyperedges connected to the *i*-th vertex. The hyperedge features aggregated by attentive hyperedge aggregation can be calculated as follows:

$$\mathbf{E}_t = \mathbf{AWH}^T \mathbf{X}_{t-1},$$

The attentive vertex aggregation module aggregates the connected hyperedge information to generate the vertex embeddings. First, the vertex features can be calculated by

$$\mathbf{X}_t^e = \mathbf{HE}_t,$$

We calculate the coefficient b_{ij} between the hyperedge i and vertex j based on the attention mechanism:

$$b_{ij} = \frac{\exp(\mathrm{LeakyReLU}(\vec{\Theta}_b \left[\overrightarrow{x_i^e} \,||\, \overrightarrow{x_j^e} \right]))}{\sum_{x_k \in N_i^e} \exp(\mathrm{LeakyReLU}(\vec{\Theta}_b \left[\overrightarrow{x_i^e} \,||\, \overrightarrow{x_k^e} \right]))},$$

where $\vec{\theta}_b \in R^{2d'}$ represents a shared attentional mechanism, parametrized by a vector, while N_i^e denotes the set of vertexes connected to the i-th hyperedge. Next, the output vertexes features $\mathbf{X}_e$ can be calculated by

$$\mathbf{X}_e = \sigma(\mathbf{BHAWH}^T \mathbf{X}_{t-1}),$$

where σ is the activation function.

Decoder

The decoder takes the low-dimensional latent embeddings of spots as input and reconstructs the spot features in the original input dimensions of the encoder. We also used hypergraph attentional networks as input to build a symmetrical decoder for encoder. After applying three layers of hypergraph attentional networks, the output of the last layer is considered as the reconstructed spot features. As for the output of t-th layer in the decoder, it could be obtained by:

$$\mathbf{X}_t = \sigma(\mathbf{X}_{t-1} \mathbf{HW}_{k-t+1}^T \mathbf{A}_{k-t+1}^T \mathbf{H}^T \mathbf{B}_{k-t+1}^T),$$

where $\mathbf{X}_0$ is the latent embeddinngs learned by the encoder, while k represents the number of layers of the encoder or decoder in our model. We set $\mathrm{k} = 3$ in this study.

Loss function

Hypergraph structured data include node features and the graph structure, both of which should be encoded by high-quality node representations. We minimized the reconstruction loss of node features as follows:

$$\sum_{i=1}^{N} \| \mathbf{X}_{in} - \mathbf{X}_{out} \|_2,$$

where $\mathbf{X}_{in}$ is the input normalized gene expression $\mathbf{X}_0$ of the encoder, and $\mathbf{X}_{out} = \mathbf{X}_k$ represents the final output of the decoder.

Downstream analysis

Clustering

To identify spatial domains based on the embeddings obtained by HyperSTAR, we employed different strategies depending on the availability of prior information about the number of labels or domains. When the number of labels is known, we utilized the mclust clustering algorithm [43]. For data without prior information about the number of domains, we used the Louvain algorithm implemented in the SCANPY package. By employing these two different clustering strategies based on the availability of prior information, we can flexibly adapt the spatial domain identification process to various datasets and research questions.

Spatial trajectory inference

To visualize the spatial trajectory inference, we employed two complementary approaches: the UMAP [44] plot and PAGA graph [45]. The UMAP plot is used to visualize the embeddings generated by the HyperSTAR. The PAGA graph is another powerful tool for visualizing spatial trajectories. PAGA is a graph-based method that combines clustering with trajectory inference, allowing for the identification of both discrete cell states and continuous transitions between them. In the context of spatial omics data, the PAGA graph can reveal the relationships between spatial domains and potential trajectories connecting them. We generated the PAGA graph using the scanpy.pl.paga function from the Scanpy package, which takes the HyperSTAR embeddings as input and produces a visually interpretable graph representation of the spatial trajectories. By combining the UMAP plot and PAGA graph, we can gain complementary and valuable insights into the spatial organization and potential trajectories present in the data, thereby facilitating the interpretation of the underlying biological processes.

SVG detection

We performed the Wilcoxon test implemented in the Scanpy package to identify spatially variable genes for spatial domains with a 1% FDR threshold (Benjamin-Hochberg adjustment).

Denoising

The denoised gene expression is generated by the decoder, and the visualization demonstrates that the denoised features are more effective and representative in distinguishing different domains.

Data availability

All datasets analyzed in this work are publicly available in raw form from their original studies. Specifically, the 10x Visium human DLPFC dataset [29] is accessible within the *spatialLIBD* package (http://spatial.libd.org/spatialLIBD). The Slide-seqV2 mouse brain dataset \cite is obtained from the Broad Institue Single Cell Portal (https://singlecell.broadinstitute.org/single_cell/study/SCP815/sensitive-spatial-genome-wide-expression-profiling-at-cellular-resolution\#study-summary). The Stereo-seq mouse brain dataset is collected from https://github.com/JinmiaoChenLab/SEDR_analyses. The Stereo-seq axolotl data is obtained from https://db.cngb.org/stomics/artista/. The breast tumor section dataset is available on 10x Genomics weibsite \url{https://www.10xgenomics.com/resources/datasets/human-breast-cancer-block-a-section-1-1-standard-1-1-0}. The osmFISH mouse brain data is collected from https://linnarssonlab.org/osmFISH/. The Xenium mouse brain data is collected from https://www.10xgenomics.com/datasets/fresh-frozen-mouse-brain-replicates-1-standard. The CODEX Barrett's esophagus data [46] is obtatined from https://datadryad.org/stash/share/1OQtxew0Unh3iAdP-ELew-ctwuPTBz6Oy8uuyxqliZk. Other data is from SODB [47]. The mouse brain reference is available in the Allen brain atlas (https://human.brain-map.org). The annotation of breast tumor section by experts is obtained by DeepST [33].

Code availability

The HyperSTAR method is implemented in Python and is available on Github: https://github.com/Ringoio/HyperSTAR.

Acknowledgements

This work has been supported by the National Key Research and Development Program of China (2022YFE0112200), the National Natural Science Foundation of China (U21A20520, 62325204), the Science and Technology Project of Guangdong Province (2022A0505050014), and the Key Area Research and Development Program of Guangzhou City (202206030009, 2023B01J1001), Australian National Health and Medical Research Council (NHMRC) Ideas

grant APP 2020646, Major and Seed Inter-Disciplinary Research Projects awarded by Monash University.

Author information

Authors and Affiliations

School of Computer Science and Engineering, South China University of Technology, Guangzhou, Guangdong, 510006, China

Yi Liao, Chong Zhang, Fei Qi, Weitian Huang & Junyu Li

Biomedicine Discovery Institute and Department of Biochemistry and Molecular Biology, Monash University, Melbourne, VIC, 3800, Australia

Zhikang Wang & Jiangning Song

Faculty of Science and Technology, Beijing Normal University-Hong Kong Baptist University United International College, Zhuhai, Guangdong, 519087, China

Shangyan Cai

Institute of Science and Technology for Brain-Inspired Intelligence, MOE Key Laboratory of Computational Neuroscience and Brain-Inspired Intelligence, MOE Frontiers Center for Brain Science, Fudan University, Shanghai, 200433, China

Center for Medical Research and Innovation, Shanghai Pudong Hospital, Fudan University Pudong Medical Center, Fudan University, Shanghai, 201399, China

Zhiyuan Yuan

Contributions

Hongmin Cai, Zhiyuan Yuan, and Jiangning Song jointly supervised the research and provided guidance throughout the study. Yi Liao, Fei Qi, and Weitian Huang conceptualized and designed the research. Yi Liao and Chong Zhang performed the research, including data collection, analysis, and interpretation. Zhikang Wang, Jinyu Li, and Shangyan Cai made significant contributions to image enhancement and manuscript language refinement. Yi Liao wrote the initial draft of the manuscript, and all authors critically reviewed and revised the manuscript for important intellectual content.

Corresponding author

Correspondence to Hongmin Cai, Zhiyuan Yuan and Jiangning Song.

Ethics declarations

Competing interests

The authors declare no competing interests.

Supplementary information

Supplementary Figs. 1-20 and Table 1.

References

1. Bressan, D., Battistoni, G., & Hannon, G. J. (2023). The dawn of spatial omics. *Science*, 381(6657), eabq4964.
2. Tang, F., Barbacioru, C., Wang, Y., Nordman, E., Lee, C., Xu, N., ... & Surani, M. A. (2009). mRNA-Seq whole-transcriptome analysis of a single cell. *Nature methods*, 6(5), 377-382.
3. Hao, Y., Hao, S., Andersen-Nissen, E., Mauck, W. M., Zheng, S., Butler, A., ... & Satija, R. (2021). Integrated analysis of multimodal single-cell data. *Cell*, 184(13), 3573-3587.
4. Efremova, M., & Teichmann, S. A. (2020). Computational methods for single-cell omics across modalities. *Nature methods*, 17(1), 14-17.
5. Rao, A., Barkley, D., França, G. S., and Yanai, I. (2021). Exploring tissue architecture using spatial transcriptomics. *Nature*, 596(7871), 211-220.
6. Moses, L., and Pachter, L. (2022). Museum of spatial transcriptomics. *Nature Methods*, 19(5), 534-546.
7. Lewis, S. M., Asselin-Labat, M. L., Nguyen, Q., Berthelet, J., Tan, X., Wimmer, V. C., ... and Naik, S. H. (2021). Spatial omics and multiplexed imaging to explore cancer biology. *Nature methods*, 18(9), 997-1012.
8. Williams, C. G., Lee, H. J., Asatsuma, T., Vento-Tormo, R., and Haque, A. (2022). An introduction to spatial transcriptomics for biomedical research. *Genome Medicine*, 14(1), 1-18.
9. Zhuang, X. (2021). Spatially resolved single-cell genomics and transcriptomics by imaging. *Nature methods*, 18(1), 18-22.
10. Larsson, L., Frisén, J., and Lundeberg, J. (2021). Spatially resolved transcriptomics adds a new dimension to genomics. *Nature methods*, 18(1), 15-18.
11. Jin, L., & Lloyd, R. V. (1997). In situ hybridization: methods and applications. *Journal of clinical laboratory analysis*, 11(1), 2-9.
12. Jensen, E. (2014). Technical review: In situ hybridization. *The Anatomical Record*, 297(8), 1349-1353.

13. Lubeck, E., Coskun, A. F., Zhiyentayev, T., Ahmad, M., and Cai, L. (2014). Single-cell in situ RNA profiling by sequential hybridization. *Nature methods*, 11(4), 360-361.

14. Eng, C. H. L., Lawson, M., Zhu, Q., Dries, R., Koulena, N., Takei, Y., ... and Cai, L. (2019). Transcriptome-scale super-resolved imaging in tissues by RNA seqFISH+. *Nature*, 568(7751), 235-239.

15. Chen, K. H., Boettiger, A. N., Moffitt, J. R., Wang, S., and Zhuang, X. (2015). Spatially resolved, highly multiplexed RNA profiling in single cells. *Science*, 348(6233), aaa6090.

16. Ke, R., Mignardi, M., Pacureanu, A., Svedlund, J., Botling, J., Wählby, C., and Nilsson, M. (2013). In situ sequencing for RNA analysis in preserved tissue and cells. *Nature methods*, 10(9), 857-860.

17. Wang, X., Allen, W. E., Wright, M. A., Sylwestrak, E. L., Samusik, N., Vesuna, S., ... & Deisseroth, K. (2018). Three-dimensional intact-tissue sequencing of single-cell transcriptional states. *Science*, 361(6400), eaat5691.

18. Metzker, M. L. (2010). Sequencing technologies—the next generation. *Nature reviews genetics*, 11(1), 31-46.

19. Goodwin, S., McPherson, J. D., & McCombie, W. R. (2016). Coming of age: ten years of next-generation sequencing technologies. *Nature Reviews Genetics*, 17(6), 333-351.

20. Mardis, E. R. (2011). A decade's perspective on DNA sequencing technology. *Nature*, 470(7333), 198-203.

21. Kanungo, T., Mount, D. M., Netanyahu, N. S., Piatko, C. D., Silverman, R., & Wu, A. Y. (2002). An efficient k-means clustering algorithm: Analysis and implementation. *IEEE transactions on pattern analysis and machine intelligence*, 24(7), 881-892.

22. Krishna, K., & Murty, M. N. (1999). Genetic K-means algorithm. *IEEE Transactions on Systems, Man, and Cybernetics*, Part B (Cybernetics), 29(3), 433-439.

23. Wolf, F. A., Angerer, P., & Theis, F. J. (2018). SCANPY: large-scale single-cell gene expression data analysis. *Genome biology*, 19, 1-5.

24. Zhao, E., Stone, M. R., Ren, X., Guenthoer, J., Smythe, K. S., Pulliam, T., ... & Gottardo, R. (2021). Spatial transcriptomics at subspot resolution with BayesSpace. *Nature biotechnology*, 39(11), 1375-1384.

25. Pham, D., Tan, X., Balderson, B., Xu, J., Grice, L. F., Yoon, S., ... & Nguyen, Q. H. (2023). Robust mapping of spatiotemporal trajectories and cell–cell interactions in healthy and diseased tissues. Nature communications, 14(1), 7739.

26. Hu, J., Li, X., Coleman, K., Schroeder, A., Ma, N., Irwin, D. J., ... & Li, M. (2021). SpaGCN: Integrating gene expression, spatial location and histology to identify spatial domains and spatially variable genes by graph convolutional network. *Nature methods*, 18(11), 1342-1351.

27. Ren, H., Walker, B. L., Cang, Z., & Nie, Q. (2022). Identifying multicellular spatiotemporal organization of cells with SpaceFlow. Nature communications, 13(1), 4076.

28. Dong, K., & Zhang, S. (2022). Deciphering spatial domains from spatially resolved transcriptomics with an adaptive graph attention auto-encoder. *Nature communications*, 13(1), 1739.

29. Maynard, K. R., Collado-Torres, L., Weber, L. M., Uytingco, C., Barry, B. K., Williams, S. R., ... & Jaffe, A. E. (2021). Transcriptome-scale spatial gene expression in the human dorsolateral prefrontal cortex. *Nature neuroscience*, 24(3), 425-436.

30. Kipf, T. N., & Welling, M. (2016). Semi-supervised classification with graph convolutional networks. arXiv preprint arXiv:1609.02907.

31. Veličković, P., Cucurull, G., Casanova, A., Romero, A., Lio, P., & Bengio, Y. (2017). Graph attention networks. arXiv preprint arXiv:1710.10903.

32. Feng, Y., You, H., Zhang, Z., Ji, R., & Gao, Y. (2019, July). Hypergraph neural networks. *In Proceedings of the AAAI conference on artificial intelligence* (Vol. 33, No. 01, pp. 3558-3565).

33. Xu, C., Jin, X., Wei, S., Wang, P., Luo, M., Xu, Z., ... & Jiang, Q. (2022). DeepST: identifying spatial domains in spatial transcriptomics by deep learning. *Nucleic Acids Research*, 50(22), e131-e131.

34. Xie, L., Dang, Y., Guo, J., Sun, X., Xie, T., Zhang, L., ... & Guo, X. (2019). High KRT8 expression independently predicts poor prognosis for lung adenocarcinoma patients. *Genes,* 10(1), 36.

35. Walker, L. C., Harris, G. C., Holloway, A. J., McKenzie, G. W., Wells, J. E., Robinson, B. A., & Morris, C. M. (2007). Cytokeratin KRT8/18 expression differentiates distinct subtypes of grade 3 invasive ductal carcinoma of the breast. *Cancer genetics and cytogenetics*, 178(2), 94-103.

36. Scott, M. K., Limaye, M., Schaffert, S., West, R., Ozawa, M. G., Chu, P., ... & Khatri, P. (2020). A multi-scale integrated analysis identifies KRT8 as a pan-cancer early biomarker. *In BIOCOMPUTING 2021: Proceedings of the Pacific Symposium* (pp. 297-308).

37. Zhang M, Eichhorn SW, Zingg B, Yao Z, Cotter K, Zeng H, Dong H, Zhuang X. Spatially resolved cell atlas of the mouse primary motor cortex by MERFISH. *Nature*. 2021 Oct 7;598(7879):137-43.

38. Codeluppi S, Borm LE, Zeisel A, La Manno G, van Lunteren JA, Svensson CI, Linnarsson S. Spatial organization of the somatosensory cortex revealed by osmFISH. *Nature methods*. 2018 Nov;15(11):932-5.

39. Wang, X., Allen, W. E., Wright, M. A., Sylwestrak, E. L., Samusik, N., Vesuna, S., ... & Deisseroth, K. (2018). Three-dimensional intact-tissue sequencing of single-cell transcriptional states. *Science*, 361(6400), eaat5691.

40. Gao, Y., Feng, Y., Ji, S., & Ji, R. (2022). HGNN+: General hypergraph neural networks. IEEE Transactions on Pattern Analysis and Machine Intelligence, 45(3), 3181-3199.

41. Chen, C., Cheng, Z., Li, Z., & Wang, M. (2020, December). Hypergraph attention networks. *In 2020 IEEE 19th International Conference on Trust, Security and Privacy in Computing and Communications* (TrustCom) (pp. 1560-1565). IEEE.

42. Veličković, P., Cucurull, G., Casanova, A., Romero, A., Lio, P., & Bengio, Y. (2017). Graph attention networks. arXiv preprint arXiv:1710.10903.

43. Fraley, C., Raftery, A. E., Murphy, T. B., & Scrucca, L. (2012). mclust version 4 for R: normal mixture modeling for model-based clustering, classification, and density estimation (Vol. 597, p. 1). *Technical report*.

44. McInnes, L., Healy, J., & Melville, J. (2018). Umap: Uniform manifold approximation and projection for dimension reduction. arXiv preprint arXiv:1802.03426.

45. Wolf, F. A., Hamey, F. K., Plass, M., Solana, J., Dahlin, J. S., Göttgens, B., ... & Theis, F. J. (2019). PAGA: graph abstraction reconciles clustering with trajectory inference through a topology preserving map of single cells. *Genome biology*, 20, 1-9.

46. Brbić, M., Cao, K., Hickey, J. W., Tan, Y., Snyder, M. P., Nolan, G. P., & Leskovec, J. (2022). Annotation of spatially resolved single-cell data with STELLAR. *Nature Methods*, 19(11), 1411-1418.

47. Yuan, Z., Pan, W., Zhao, X., Zhao, F., Xu, Z., Li, X., ... & Yao, J. (2023). SODB facilitates comprehensive exploration of spatial omics data. *Nature Methods*, 20(3), 387-399.